연합학습 모델에 대한 특성 추론 공격 및 방어 기법에 대한 연구

김현준[1], 조윤기[1], 백윤흥[1]
[1]서울대학교 전기정보공학부, 반도체공동연구소
hjkim@sor.snu.ac.kr, ygcho@sor.snu.ac.kr, ypaek@sor.snu.ac.kr

A Survey on Property Inference Attack and Defense Technique for Federated Learning Model

Hyun-Jun Kim[1], Yun-Gi Cho[1], Yun-Heung Paek[1]
[1]Dept. of Electrical and Computer Engineering and Inter-University Semiconductor Research Center (ISRC), Seoul National University

요 약

본 논문에서는 연합학습 모델을 타겟으로 하는 특성 추론 공격 및 방어 기법과 관련된 연구들을 소개한다. 연합학습 시스템에 특화된 2가지 특성 추론 공격 및 이에 대한 방어 기법들에 대해 정리하고, 향후 연구 방향을 조망하고자 한다.

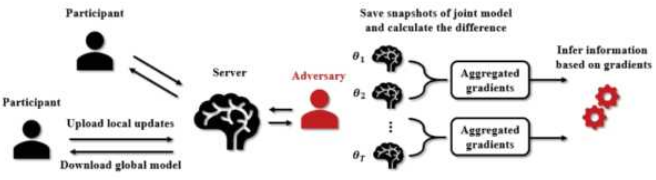
1. 서론

연합학습 (federated learning, FL)[1]은 전역 모델 (global model)을 학습하는 서버와 국소 모델 (local model) 학습을 개별로 진행하는 클라이언트들로 구성되어 있으며, 각 클라이언트의 개인 데이터셋에 대한 직접적인 액세스 없이 전체 데이터셋에 대해 AI 모델을 학습시킬 수 있다는 장점이 있다. AI 모델이 발전하면서 모델의 복잡성이 향상됨에 따라 다량의 학습 데이터셋이 요구되었고, 이에 따라 연합학습 모델에 대한 수요도 증가하였다. 특히 의료 데이터[2], 금융 데이터[3] 등 민감한 개인 데이터를 다루는 분야에서 주로 연합학습과 관련된 연구가 많이 진행되었다. 해당 분야에서는 개인 프라이버시 (privacy)가 보장되는 것이 중요하며, 보안 분야에서는 개인 프라이버시를 해칠 수 있는 연합학습 모델의 취약성에 대한 많은 연구를 진행해왔다.

최근 개발된 추론 공격들은 데이터셋을 정확히 알지 못하더라도 연합학습에 사용되는 업데이트, 모델 등에서 클라이언트의 데이터셋에 대한 여러 정보들을 추출해낸다. 특히 특성 추론 공격 (property inference attack)은 모델의 메인 태스크에서 다루는 클래스와 관련이 없으면서도 입력 이미지에 포함된 특성에 대한 정보를 추출하는 공격이다. 예를 들면 어떤 모델이 사람의 얼굴 이미지를 입력으로 받아서 나이를 추론하는 것이 메인 태스크라고 한다면, 해당 사람이 안경을 썼는지 안 썼는지와 같은 정보를 추출해낼 수 있다. 이러한 정보들은 공격자에 의해 특정 특성을 가진 개인을 식별해내는 데에 악용될 수 있다. 본 논문에서는 연합학습 시스템에 특화된 특성 추론 공격들 및 방어 기법들을 알아보고, 이를 기반으로 향후 관련 연구 방향을 탐색하고자 한다.

2. 연합학습 모델에 대한 특성 추론 공격

2-1. Exploiting Unintended Feature Leakage in Collaborative Learning [4]



(그림 1) 특성 추론 공격 개요도

해당 연구에서는 공격자가 연합학습에서 하나의 클라이언트로 참여하여 특정 피해자 (victim) 클라이언트가 학습에 사용하는 특성을 추론해내는 공격을 다룬다. 공격자는 매 라운드마다 전역 모델 패러미터를 저장해두고, 현재 라운드 전역 모델 패러미터에서 이전 라운드 전역 모델 패러미터를 빼서 이전 라운드에서 클라이언트들이 서버에 보낸 업데이트

값을 평균낸 합계 그래디언트 (aggregated gradient)를 계산한다. 만약 해당 라운드에 참여한 클라이언트 중 하나 이상의 클라이언트가 학습에 사용한 데이터 중 특정 특성을 가진 이미지가 포함된다면 합계 그래디언트에 해당 이미지에 대한 패턴이 나타난다. 공격자는 전역 모델과 자기 자신이 가진 예비 데이터셋 (모델이 학습된 데이터셋과 비슷한 분포를 가진 추가 데이터셋)을 활용하여 가상 연합학습 모델을 학습시키고 여기서 나타난 합계 그래디언트 패턴을 랜덤 포레스트 모델에 학습한다. 그리고 해당 랜덤 포레스트 모델에 실제 라운드에서 계산된 합계 그래디언트 패턴을 입력으로 넣으면 해당 라운드에서 특정 특성을 가진 이미지가 학습에 사용되었다는 정보를 알아낼 수 있다. 추가로 해당 연구에서는 단순히 정보를 알아내는 것에서 더 나아가 능동적인 특성 추론 공격 방식도 제안하고 있다. 공격자는 추가 분류기 (classifier)를 자신의 국소 모델에 추가하여 전역 모델이 특정 특성을 가진 이미지가 학습에 사용되었는지 안 사용되었는지를 더 잘 식별할 수 있도록 학습을 수행한다. 이렇게 간접적으로 조작된 전역 모델은 좀 더 특성 추론 공격에 취약하게 된다.

2-2. Property Inference from Poisoning [5]

해당 연구는 전역 모델에 대해 포이즈닝 (poisoning) 공격을 수행해서 특성 추론 공격에 더 취약하게 만드는 기법을 다루고 있다. 공격자는 먼저 변형된 학습 데이터를 주입해서 전역 모델을 오염시키고, 일련의 쿼리 (query)들을 서버에 보내서 추론 결과들(라벨로 주어짐)을 받는다. 그리고 이 결과들을 활용해서 학습 데이터에 대해 어떤 특성의 평균값(학습 데이터 내 포함 비율)이 특정 쓰레스홀드 (threshold) 값보다 더 높은지 낮은지를 알아내는 것이 목적이다. 해당 공격은 라벨이 0인지 1인지 애매한 데이터 포인트들을 먼저 식별해내고 해당 데이터 포인트들 중 특정 특성 값이 1인 데이터들의 라벨을 전부 0 혹은 전부 1로 통일시켜서 조작해서 전역 모델 학습에 참여한다. 공격자가 애매한 데이터 포인트들을 쿼리로 서버에 보내면 전역 모델 학습에 사용된 데이터 셋에 포함된 특정 특성을 가진 데이터 포인트들의 비율을 간접적으로 알아낼 수 있다.

3. 연합학습 모델에 대한 특성 추론 공격 방어 기법

특성 추론 공격을 포함한 추론 공격들에 대한 방어 기법으로는 첫 번째로 정규화 (regularizer) 기법이 있다. L2 손실 함수, 드롭아웃 (dropout)과 같이 AI 모델의 오버피팅을 방지하는 기법들은 클라이언트에서 서버로 보내는 업데이트의 일부 정보를 제거하기 때문에 결과적으로 학습 데이터셋에 대한 정보도 일부 제거하여 추론 공격의 성능이 낮아지게 된다. 두 번째 기법은 차등 프라이버시 (differential privacy, DP)가 있다. 해당 기법은 업데이트에 노이즈를 추가하여 학습 데이터셋에 대한 정보에도 노이즈가 가해지고 공격을 방해할 수 있다. 하지만 앞서 소개한 연구들 [4], [5] 및 특성 추론 공격에 대한 차등 프라이버시 기법을 테스트한 연구 [6]에서는 위와 같은 방어 기법들이 유효하지 않거나 방어는 되지만 모델 성능을 많이 떨어뜨리는 문제점이 있어 실제로 적용할 수 없다고 보고하였다.

4. 차후 연구 방향

차후에는 위와 같은 특성 추론 공격들을 사전 방지하거나 정보가 유출되더라도 가짜 정보를 유출시키는 등 방어 기법에 대한 추가 연구가 필요하다. 한 가지 유효한 방법으로는 연합학습 도중 전역 모델 학습 단계에서 자체적으로 서버 내에서 특성 추론 공격들을 수행해보고 만약 유출된 정보가 실제 정보와 동일하다면 모델의 성능을 적게 떨어뜨리면서도 해당 정보만 다른 값으로 조정할 수 있도록 업데이트 방식을 조정하거나 차등 프라이버시 노이즈 값을 조정할 수 있을 것으로 예상된다.

5. 결론

본 논문에서는 연합학습 모델을 대상으로 하는 특성 추론 공격 및 방어 연구들의 동향을 살펴보았고, 각 연구에서 사용한 기법들에 대해 자세히 알아보았다. 아직 방어 기법 측면에서 발전이 더 필요한 상황이며 개선이 가능할 것으로 기대된다.

6. Acknowledgement

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원 (No.2020-0-01840,스마트폰의 내부데이터 접근 및 보호 기술 분석)과 2023년도 BK21 FOUR 정보기술 미래인재 교육연구단의 지원을 받았으며, 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 지원을 받아 수행된 연구임 (IITP-2023-2020-0-01602).

참고문헌

[1] Konečný, Jakub, et al. "Federated learning: Strategies for improving communication efficiency." arXiv preprint arXiv:1610.05492. 2016.

[2] Antunes, Rodolfo Stoffel, et al. "Federated learning for healthcare: Systematic review and architecture proposal." ACM Transactions on Intelligent Systems and Technology (TIST) 13.4 pp.1-23. 2022.

[3] Long, Guodong, et al. "Federated learning for open banking." Federated Learning: Privacy and Incentive. Cham: Springer International Publishing, pp.240-254. 2020.

[4] Melis, Luca, et al. "Exploiting unintended feature leakage in collaborative learning." 2019 IEEE symposium on security and privacy (SP). IEEE, United States, 2019.

[5] Mahloujifar, Saeed, Esha Ghosh, and Melissa Chase. "Property inference from poisoning." 2022 IEEE Symposium on Security and Privacy (SP). IEEE, United States, 2022.

[6] Naseri, Mohammad, Jamie Hayes, and Emiliano De Cristofaro. "Local and Central Differential Privacy for Robustness and Privacy in Federated Learning." NDSS, United States, 2022.